Malaria Life Cycle Intensifies Both Natural Selection and Random Genetic Drift

Classification: Biological Sciences

Title: The malaria life cycle intensifies both natural selection and random genetic drift.

Short title for mobile devices and RSS feeds: Selection and drift in the malaria life cycle

Authors: Hsiao-Han Chang[a], Eli L. Moss[b], Daniel J. Park[b], Daouda Ndiaye[c], Soulyemane Mboup[c], Sarah K. Volkman[b,d,e], Pardis C. Sabeti[a,b,d], Dyann F. Wirth[b,d], Daniel E. Neafsey[b], and Daniel L. Hartl[a]

Affiliations:

[a] Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138;

[b] Broad Institute of MIT and Harvard, Cambridge, MA 02142;

[c] Faculty of Medicine and Pharmacy, Université Cheikh Anta Diop de Dakar, BP 5005, Dakar Fann, Sénégal;

[d] Department of Imunology and Infectious Diseases, Harvard School of Public Health, Boston, MA 02115;

[e] School for Nursing and Health Sciences, Simmons College, Boston, MA 02115

Corresponding author:

Hsiao-Han Chang

BioLab 2105, 16 Divinity Avenue, Cambridge, MA 02138

Phone: 617-642-7897

e-mail: hhchang@hsph.harvard.edu

Keywords: *Plasmodium*, life cycle, genetic drift, selection efficiency

Author contributions:

H.-H.C. and D.L.H. designed research; H.-H.C. performed research; D.E.N., D.F.W., P.C.S., S.K.V., R.C.W., D.N. and S.M. contributed reagents/materials/analysis tools; E.L.M., and D.J.P. processed the raw sequence data; H.-H.C. analyzed data; and H.-H.C and D.L.H. wrote the paper.

Abbreviations: WF, Wright-Fisher

Significance: Genomic sequences of 159 isolates of the malaria parasite *Plasmodium falciparum* exhibited highly unusual patterns of single-nucleotide polymorphism. We hypothesized that these patterns might result from the repeated bottlenecks in host–vector and vector–host transmission as well as the intense competition between parasites within a single host. Computer simulations of the malaria life cycle recapitulated the unusual patterns of polymorphism observed. In the classical Wright-Fisher (WF) model in population genetics, random changes in gene frequency caused by finite population size (random drift) diminish the efficiency of natural selection. The trade-off between drift and selection has been widely assumed to be robust to details of the life cycle. In the malaria parasite, however, both selection and drift are simultaneously enhanced.

Abstract

Analysis of genome sequences of 159 isolates of *Plasmodium falciparum* from Senegal yields an extraordinarily high proportion (26.85%) of protein-coding genes with the ratio of nonsynonymous to synonymous polymorphism greater than one. This proportion is much greater than observed in other organisms. Also unusual is that the site-frequency spectra of synonymous and nonsynonymous polymorphisms are virtually indistinguishable. We hypothesized that the complicated life cycle of malaria parasites might lead to qualitatively different population genetics from that predicted from the classical Wright-Fisher (WF) model, which assumes a single random-mating population with a finite and constant population size in an organism with nonoverlapping generations. This paper summarizes simulation studies of random genetic drift and selection in malaria parasites that takes into account their unusual life history. Our results show that random genetic drift in the malaria life cycle is more pronounced than under the WF model. Paradoxically, the efficiency of purifying selection in the malaria life cycle is also greater than under WF, while the relative efficiency of positive selection varies according to conditions. Additionally, the site-frequency spectrum under neutrality is also more skewed toward low frequency alleles than expected with WF. These results highlight the importance of considering the malaria life cycle when applying existing population genetic tools based on the WF model. The same caveat applies to other species with similarly complex life cycles.

\body

Introduction

Malaria, caused by the parasite, *Plasmodium falciparum*, is one of the major causes of death worldwide. To aid the development of vaccines and drug treatments for malaria, researchers have studied the *P. falciparum* genome and identified genes that are essential to malaria parasites as well as genes that are related to drug-resistance phenotypes using population genetic tools (1-6). Researchers have also focused on particular genes related to drug resistance and characterized the evolutionary pathways of emerging drug resistance using *Escherichia coli* and *Saccharomyces cerevisiae* as model systems (7-10).

Malaria parasites have a complex life cycle with two types of host organisms — humans and female *Anopheles* mosquitoes. Malaria parasites are transmitted from mosquito to humans through the bite of an infected mosquito. In the human host, the parasite reproduces asexually multiple times, and the within-human population size increases from 10–10^2 at the time of infection to 10^8–10^{13} within a few weeks. When another female mosquito feeds on the blood of the infected human, 10–10^3 malaria gametocytes are transmitted back to the mosquito host, and these immature gametes undergo maturation, fuse to form zygotes, undergo sexual recombination and meiosis, and the resulting haploid cells reproduce asexually and form sporozooites that migrate to the salivary glands to complete the life cycle (11). These features of the malaria life cycle

pose potential problems when attempting to analyze population genetic data using simpler models of life history and reproduction.

Much of population genetics theory is based on the concept of a Wright-Fisher (WF) population (12, 13). In the WF model, the population size is constant, generations are non-overlapping, and each new generation is formed by sampling parents with replacement from the current generation. The major differences between the malaria life cycle and the WF model are that each malaria life cycle includes two transmissions, multiple generations of asexual reproduction, and population expansions and bottlenecks. Before population genetic inferences can be conducted through analysis based on WF assumptions, it is necessary to determine whether the malaria life cycle is sufficiently well described by the WF model. If the life cycle impacts features of population genetics, then inferences based on conventional interpretations of the WF model may need to be adjusted.

In a previous study based on only 25 parasite isolates, we observed two unusual patterns in the *P. falciparum* genome that had not been reported in any other organism (4). First, we observed synonymous and nonsynonymous site-frequency spectra that were more similar than expected given that nonsynonymous sites likely experience stronger selection. Second, almost 20% of the genes showed a ratio of nonsynonymous to synonymous polymorphism (π_N/π_S) greater than 1. In *D. melanogaster* (14), fewer than 2% of the genes have π_N/π_S greater than 1. Because nonsynonymous mutations result in changes to

amino acids, they are likely to have a deleterious effect and exist in low frequencies in the population or be completely eliminated. In other organisms, the nonsynonymous site-frequency spectrum is more skewed toward low-frequency alleles than the synonymous site-frequency spectrum; examples include humans (15-17), *Oryctolagus cuniculus* (18), *Drosophila melanogaster* (19), and *Capsella grandiflora* (20).

Potential explanations for these unusual patterns including sequencing error and annotation error could be ruled out, and dramatically relaxed or diversifying selection for almost 20% of protein-coding genes seems unlikely. Although selection on antigens could possibly explain the high prevalence of genes with π_N greater than π_S, the nonsynonymous site-frequency spectrum is skewed toward low frequency alleles, which is not what one would expect if frequency-dependent balancing selection explains the phenomenon. Because of the complexities of the malaria life cycle, we wondered whether the malaria life cycle itself could explain part of these unusual patterns. More recent work in *P. vivax*, a close relative with similar life history to *P. falciparum*, also revealed large numbers of genes with π_N/π_S greater than 1 (21), supporting the idea that factors common to *Plasmodium* species but different from most other species may cause allele-frequency patterns that deviate from WF expectations.

Although the behavior of the WF model is relatively robust to deviations from many underlying assumptions, there are examples in which the WF model is known to perform poorly. For instance, it was recently shown that the effect of

selection is increased relative to the WF model when the distribution of offspring number allows occasional large family sizes (22). While Otto and Whitlock (1997) define a “fixation effective population size," they also emphasize that it is a function of the selection coefficient when population size changes in time (23). Their results highlight the importance of studying the effect of various reproductive mechanisms on basic evolutionary outcomes. Although there has been research on the evolution of drug resistance in malaria parasites, in both mathematical models and computational simulations (24-27), it has not been ascertained whether the underlying processes of random genetic drift, natural selection, and their interactions yield outcomes in the malaria life cycle that are congruent with those of the WF model.

Here, we sequenced 159 genomes of *P. falciparum* isolates from Senegal and studied the patterns of polymorphism. We find virtually identical site-frequency spectra for synonymous and nonsynonymous polymorphisms, and 26.85% of the protein-coding genes exhibit $\pi_N/\pi_S > 1$. To investigate whether the life cycle could explain the observed unusual patterns of polymorphism, we used Monte-Carlo simulations to examine how the malaria life cycle influences random genetic drift, natural selection, and their interactions. First, we compared quantities from generation to generation between a malaria model and the WF model, including the number of mutant alleles after one generation and probability of loss. Second, we considered properties on a longer time scale, including time to fixation or loss, segregation time, and probability of fixation or

loss. Third, we simulated the site-frequency spectrum under a neutral model with the malaria life cycle. The flexibility of the simulation framework enables us to investigate various combinations of selection coefficients. Finally, we discuss the simulation results and suggest how the malaria life cycle could possibly lead to these unusual population genetic patterns.

Results

Unusual patterns of genetic diversity

Among genome sequences of 159 isolates of *P. falciparum* from Senegal, we calculated the ratio of nonsynonymous to synonymous polymorphism (π_N/π_S) for genes with synonymous polymorphism π_S greater than zero. Among 4395 such genes, 1157 (26.85% ± 0.67%) exhibited a ratio π_N/π_S greater than 1. We also compared the synonymous and nonsynonymous site-frequency spectra, and found that they are indistinguishable (Mann-Whitney test, *P* value = 0.46) (Fig. 1). These results are consistent with an earlier report based on a much smaller sample size (4).

Allele frequency change from generation to generation

The forward-time Monte-Carlo simulation framework of the malaria life cycle is described in detail below (Fig. 2). The key parameters in the model are: the selective advantage or disadvantage of a mutant allele within the human host

per cycle of asexual reproduction (s_h), the selective advantage or disadvantage of a mutant allele within the mosquito vector per cycle of asexual reproduction (s_m), the transmission advantage or disadvantage of a mutant allele from the human host to the mosquito vector (t_m) and from the mosquito vector to the human host (t_h), the number of human hosts (N), the number of mosquito vectors per human host (a), the number of sporozoites and gametocytes transmitted between the vector and the human host (D), the probability that a parasite undergoes replication in a given asexual cycle (P), and the number of asexual generations that the parasite population remains at its maximum size (i.e. peak parasitemia) in the human host (e). Using this model, we examined random genetic drift during the malaria life cycle by comparing the probability of loss of a selectively neutral mutant allele after one complete life cycle (regarded as one generation in the malaria model) with that after one generation in the WF model. In the malaria model, the probability of loss of a new neutral allele is as high as 74%, whereas it is approximately $e^{-1} \approx 37\%$ under the WF model (e^{-1} is the probability of observing 0 outcomes in a Poisson model with mean 1, which is the approximation of a binomial model with large n and small p where $np = 1$; the latter is equivalent to the WF model*).* Moreover, while the average frequency is the same, the variance in the frequency of the mutation after one generation in the malaria model is higher than that in the WF model [5.71 (malaria) vs 1.00 (WF)]. These discrepancies indicate that random genetic drift has much stronger effects in the malaria model.

In the case of a non-neutral allele, the probability of loss in the malaria model is greater than that in the WF model (Fig. 3A), irrespective of whether the mutation is beneficial or deleterious. Interestingly, the average number of copies of a mutant allele one generation after its occurrence is also more extreme in the malaria life cycle (Fig. 3B). On average, after one generation, beneficial alleles leave more copies than in the WF model, and deleterious alleles leave fewer copies, suggesting that selection works more efficiently in the malaria life cycle. This result is in contrast to that expected from existing population genetic theory. It is commonly thought that, when random genetic drift increases, the selection efficiency must decrease (28). The results in Fig. 3 imply that, in the malaria life cycle, random genetic drift and selection efficiency can increase simultaneously.

We tested whether the difference between the malaria life cycle and the WF model is sensitive to other parameters by varying the values of other parameters in the simulation including a, e, P and D. The results show differences in detail, but are qualitatively consistent (Fig. S1).

Allele frequency change on a longer time scale

We then considered properties on a longer time scale, including the segregation time (the average time until a mutation becomes fixed or lost in the population), the time to fixation, time to loss, and the fixation probability. We treated one complete life cycle as one generation in the malaria model and compared it with one generation in the WF model. The results indicate that the

segregation time in the malaria model is shorter than in the WF model (Fig. 4A). The mutations segregate on average for less than 8 generations in the malaria model, even when the selection coefficient is as high as 0.1, because of the enhanced genetic drift during the malaria life cycle. The shortening of the segregation time also indicates that a large proportion of segregating sites in the genome of malaria parasites are likely to be recently derived.

The time to fixation for beneficial mutations in the malaria model is shorter than that in the WF model when the selection coefficient is smaller than a threshold value, and longer for larger selection coefficients (Fig. 4B). When the selection coefficient is small, the time to fixation of beneficial mutations is shorter in the malaria model because, after a mutation becomes fixed in the population of parasites infecting one host, it benefits from a greater increment in allele frequency in each generation because of the transmission of multiple parasites between human and mosquito. However, in the simulation shown in Fig. 4B, because only the within-host selection coefficient (s_h or s_m) is positive and the transmission coefficient (t_M or t_H) is 0, there is no transmission advantage between hosts for a beneficial allele and this could lower the fixation time. Thus, when the selection coefficient exceeds a threshold, selection in the WF model is so efficient that fixation takes less time than in the malaria model in spite of the transmission of multiple parasites. Nevertheless, the probability of fixation of beneficial alleles in the malaria model is always smaller than that in the WF

model (Fig. 4C) owing to the enhanced random genetic drift and the stochastic nature of parasite transmission among hosts.

The time to loss for deleterious mutations in the malaria model is also shorter than that in the WF model (Fig. 4D), suggesting that purifying selection is more efficient in the malaria model and deleterious mutations are removed from the population very quickly, hence segregating mutations in the malaria parasite are less likely to be deleterious than mutations observed in other organisms with similar effective population sizes that evolve in accord with the WF model.

We examined whether these results are sensitive to values of parameters other than the selection coefficient by varying the values of the parameters *a*, *e*, *P* and *D* in the simulation. The results are again qualitatively consistent and differ only quantitatively (Fig. S2).

Preferential transmission

It has been suggested that genetic factors influence the rate of conversion of gametocytes into male or female gametes (29). Because gametocyte differentiation is critical for forming zygotes in the mosquito host and successful transmission, transmission between hosts could be affected by mutations in the parasite genome. We therefore performed the simulations in which transmission probabilities could be altered by mutations.

When mutation only affects the transmission probability (t_m-only model in Fig. 5A), beneficial mutations have even shorter segregation times than when

selection occurs only in the host (s_h-only model), and deleterious mutations have slightly longer segregation times than in the s_h-only model (Fig. 5A). Fixation times for beneficial mutations in the t_m-only model and the $t_m = s_h$ model are both shorter than in either the s_h-only model or the WF model (Fig. 5B), suggesting that transmission advantage or disadvantage is major determinant of the fixation time. The fixation probabilities for beneficial mutations in the t_m-only model and the $t_m = s_h$ model are larger than in the s_h-only model, but still smaller than in the WF model (Fig. 5C). Among the three malaria models, the $t_m = s_h$ model has the greatest efficiency for positive selection because it has higher probability of fixation and shortest time to fixation for beneficial alleles. All three malaria models show patterns that are qualitatively different from the WF model.

Site-frequency spectrum

We also simulated the site-frequency spectrum for neutral alleles when the sample size is 159, matching the number of genomes sequenced. The result shows that the malaria life cycle skews the site-frequency spectrum to the lower frequency alleles (Fig. 6). When interpreted in the WF framework, this skewing implies an increasing parasite population size. But in the simulations the parasite and host population sizes do not change; the skewing is entirely the result of the differences between the actual malaria life cycle and that assumed in the WF model. The difference is in part due to the population expansion within hosts in each generation; this makes estimation of parasite demographic history more

difficult than in other organisms, and a previous study may overestimate the population expansion (4). The simulation results also imply that intrinsic differences in evolutionary processes caused by the complex malaria life cycle alter the null distributions of tests of selection based on the site frequency spectrum (30-32). This emphasizes the importance of considering the complexities of the malaria life cycle when analyzing genomic data to infer demographic history and to identify genes under selection.

Discussion

We examined complete genome sequences of 159 malaria parasites from Senegal and observed that extraordinarily high proportion of genes (26.85% of 4395 protein-coding genes) with π_N/π_S ratio greater than 1. We also observed that the site-frequency spectrum of polymorphisms was indistinguishable between synonymous and nonsynonymous sites in protein-coding genes. Our simulations demonstrate that both of these unexpected features in the data could result from the complex life cycle of malaria and its effects on allele-frequency change. In comparing the malaria life cycle with the classical WF model, we found that mutations in the parasite population segregate for a shorter time (Fig. 4A) and that, for deleterious alleles, the probability of loss is greater (Fig. 3B) and the time to loss shorter (Fig. 4D). These results suggest that purifying selection works more efficiently in the malaria life cycle. The malaria parasite shows evidence for efficient selection that affects base composition at synonymous sites and in

intergenic regions, supporting the inference that purifying selection is efficient. For example, there are significant differences between the C/G to A/T and the A/T to C/G site-frequency spectra in the genomes of malaria parasites from Senegal (4). Because purifying selection works so efficiently in the parasite life cycle, and the probability of loss is so high, most of the segregating mutations are either very new or very nearly neutral. A corollary result is that the expected difference between the synonymous and nonsynonymous site-frequency spectra is reduced.

In addition, because of the high efficiency of purifying selection in the parasite, sites with relatively small selection coefficients could nevertheless have their ultimate fate determined by selection whereas the same selection coefficients would segregate as nearly neutral in the WF model. This finding suggests that polymorphisms at synonymous sites, which are commonly thought to be effectively neutral or under weak selection, experience more efficient selection in the malaria parasite. Although both nonsynonymous and synonymous polymorphisms are expected to be reduced in frequency due to the enhanced efficiency of selection, one expects a greater effect on synonymous sites than on nonsynonymous sites because an increased efficiency of selection will have a greater effect on slightly deleterious alleles than on strongly deleterious alleles (Fig. S3). The upshot of a greater reduction in π_S than in π_N is an increased proportion of genes with $\pi_N/\pi_S > 1$. It should also be noted that *P. vivax*, which has a life cycle that is similar to that of *P. falciparum*, also shows

a large number of genes with π_N/π_S greater than 1 (21). The observed population genetic patterns are not likely to be due to the structure of the parasite population being divided between hosts. In a structured population with limited gene flow, the coalescence time and the level of polymorphism are expected to be higher than a random-mating population. However, because the structured population affects both synonymous and nonsynonymous sites, it does not increase π_N/π_S or decrease the expected difference between synonymous and nonsynonymous site-frequency spectra due to different levels of selective constraints.

In regard to positive selection, whether the efficiency of positive selection is higher or lower with the malaria life cycle depends on the selection coefficient. When the selection coefficient is small, the time to fixation for beneficial alleles is shorter than in the WF model. When a mutation increases both the selection coefficient and the transmission probability, or when a mutation increases only the transmission probability, the time to fixation in the malaria life cycle is less than in the WF model across the whole range of simulated selection coefficients. However, it should be noted that the probability of fixation in the malaria model is less than in the WF model owing to the enhanced random genetic drift. Otto and Whitlock (1997) have studied the probability of fixation of beneficial alleles in a model with cyclical changes in population size, and they emphasize the importance of the cycle time relative to the time scale of selection. While their model yields important insights (*e.g.*, the probability of fixation of a mutant allele depends on when in the population cycle the mutation occurs), the model does

346 not really apply to malaria owing to the natural population subdivision among
347 hosts. In malaria, each infected host, and each infected vector, represents a
348 separate local population or deme of parasites. The situation is further
349 complicated by the fact that the selection coefficient may depend on events that
350 take place in the host, in the host-vector transmission, in the vector, in the vector-
351 host transmission, or in any combination of these stages.

352 Looking at the larger picture, random genetic drift and natural selection are
353 two major forces that shape genetic variation. In standard population genetic
354 models including the WF model, when population size increases, the influence of
355 random genetic drift decreases and that of natural selection increases, and
356 conversely (28). Our main finding reported in this paper is that, in the malaria life
357 cycle, both random genetic drift and natural selection are intensified
358 simultaneously. Because of the unique parasite features of multiple asexual
359 generations, population expansion within hosts, and stochastic transmission in
360 each iteration of the life cycle, natural selection and random genetic drift can both
361 increase at the same time. The lack of any tradeoff between random drift and
362 selection contrasts with classical theoretical population genetics, and it
363 demonstrates the importance of taking the parasite life cycle into account when
364 interpreting genomic sequence data. Many microbial populations and parasite
365 species have population growth and population bottlenecks during their life
366 cycles, and these species may be affected by similar dynamics. Our results also
367 suggest that other parasite species that are transmitted among hosts with

population-size expansion within hosts may evolve in a way that is qualitatively different from the Wright-Fisher model. Caution is therefore in order when interpreting data based on standard population genetic methods in organisms with unconventional life histories.

Recombination is another important force in shaping genomic variation (33-36). In the malaria life cycle, mutation can happen in any asexual generation within host or vector, but meiotic recombination takes place only once per life cycle within the mosquito host. As well, the proportion of multiple infections differs among geographical locations, and therefore the level of inbreeding differs. Previous theoretical work has reported the effect of hitchhiking and partial selfing on genetic variation (37), modeled the hitchhiking effect of drug-resistant alleles, and demonstrated that selection and recombination cannot be decoupled in the malaria life cycle (27). While the observed level of linkage disequilibrium in the Senegal population is very low (4), the next step in the modeling will be to allow for the possibility of mixed infections in a multilocus model to examine the relation between transmission intensity and linkage disequilibrium.

Comparison of malaria with WF across multiple generations

It could be argued that the simulated malaria life cycle shows different effects of drift and selection from the WF model simply because we treat multiple asexual generations in one life cycle as one "generation." We therefore compared the probability of loss and the average number of mutations after

multiple WF generations with the model of the malaria life cycle (Fig. S4). The comparison makes it clear that, even if we increase the generation numbers in the WF model, the malaria life cycle gives qualitatively different results from the WF model. The change in probability of loss as a function of change in selection coefficient is consistently greater in the malaria life cycle.

The finding that the discrepancies between malaria and the WF model cannot simply be remedied by redefining a "generation" in malaria is somewhat reassuring. Malaria researchers traditionally regard a single malaria "generation" as the time between sexual cycles. If the number of asexual cell divisions between sexual cycles also had to be taken into account, then certain oddities would arise. For example, applying the same logic to humans would produce a generation time in males that is substantially longer than that in females owing to the larger number of germ-cell divisions in males.

An "effective *s*" for malaria?

The effective population size of a real population is defined as the population size of an ideal population that has the same level of random genetic drift as the population of interest (12). This concept is useful when the population under consideration deviates in specified ways from the WF model. In principle, one could try to define an "effective selection coefficient" in malaria and use this to make predictions or inferences from tools based on the WF model. The effective selection coefficient for malaria would correspond to that in an ideal WF

412 model that has the same population dynamics as in the malaria model. To
413 evaluate the feasibility of this approach, we identified selection coefficients in the
414 WF model that have the same probability of loss and average number of
415 mutations as in the malaria model (Table S1). We found that it is not possible to
416 fit these two properties at the same time, and hence an effective selection
417 coefficient that holds for all aspects of the population dynamics does not exist.
418

419 Comparison with selectively neutral conditions

420 Besides comparing the absolute values of time to fixation or probabilities
421 of fixation in the two models, we also examined the ratio of these quantities in the
422 neutral case in the malaria model versus the WF model (Fig. S5). Fig. S5 shows
423 that both the relative time to fixation of beneficial alleles and the relative time to
424 loss of deleterious alleles are longer in the malaria model. This difference
425 indicates that equal transmission probability among hosts reduces the fold-
426 difference between the neutral and selective cases. However, for deleterious
427 mutations, the absolute time to loss is probably more relevant than the relative
428 values because most of deleterious mutations are lost within few asexual
429 generations in the host in which the mutation happens, and therefore neutral
430 transmission among hosts does not play an important role. Random genetic drift
431 (and the probability of loss) is so high that even neutral mutations are quickly lost,
432 and the fold-difference between deleterious mutations and neutral mutations is

smaller than in the WF model. Hence, this result also supports the conclusion of greater efficiency of purifying selection during the malaria life cycle.

In summary, this study used computer simulation to investigate the effect of the malaria life cycle on population genetic behaviors. The results suggest that both genetic drift and the efficiency of purifying selection are intensified by the malaria life cycle. Because these two properties typically cannot be enhanced at the same time in traditional models, this demonstrates the intrinsic differences between the WF model and the malaria life cycle. Furthermore, the site-frequency spectrum in the malaria model is more skewed toward low frequency alleles even if the host population size remains constant. Our study suggests that malaria life cycle itself leads to unusual patterns of polymorphism, and hence life cycle should be considered explicitly in order to study the evolution of malaria parasites or other organisms with similar life cycle through patterns of genetic diversity.

Materials and Methods

Dataset and data processing

A total of 159 isolates of *P. falciparum* from Senegal were sequenced and investigated in this study. Sample preparation method for each isolate is listed in Table S2. Genomic DNA was sequenced using Illumina HiSeq machines and sequence reads were aligned to the *P. falciparum* 3D7 reference genome (38)

455 from PlasmoDB version 9.0 (http://PlasmoDB.org) using Burrows-Wheeler
456 Aligner (BWA) 0.6.2 (39) and the SAMTools version 0.1.18 (40). Genotypes were
457 called from the reads for each isolate separately using the GATK Unified
458 Genotyper version 2.1-13-g1706365 (diploid mode with hard-filtering of
459 heterozygotes) (41) because calling genotypes jointly calls more heterozygous
460 calls in error. Picard version 1.473 was used to strip preexisting alignment
461 annotations from BAM files prior to realigning against our chosen reference
462 sequence, the PlasmoDB v9.0 3D7 assembly. Sites with PHRED -style GQ
463 scores above 30 and QUAL scores above 60 were kept. Repeat-rich sequences
464 near the telomeres of each chromosome were excluded from the analyses (Table
465 S3) (42). Highly variable *PfEMP1* (*var*) genes were excluded from the analyses
466 because the reads from these genes are difficult to align to the reference genome
467 correctly. Sequences were submitted to the NCBI Sequence Read Archives (SRA)
468 under the accession numbers SRP000316, SRP000493, SRP003502,
469 SRP007838, SRP007883, SRP007923, and SRP012397.
470

471 Sequence analysis

472 Nonsynonymous and synonymous polymorphism (π_N and π_S) were
473 calculated using the same method as described (4). For the analyses of
474 synonymous and nonsynonymous site-frequency spectra and polymorphism, we
475 only used codons where each of the three nucleotides has less than 20% missing
476 data and less than two nucleotides are polymorphic.

Simulation

To simulate the evolution that takes into account the malaria life cycle, we used the following forward-time Monte-Carlo simulation framework (Fig. 2):

(i) Assume there are N human hosts and $a \times N$ mosquito hosts, with D parasites (sporozoites) transmitted from the mosquito host to the human host. The initial condition is the number of mutations in the initial parasite pool within each human host.

(ii) Within a human host, the probability of a parasite that carries a particular allele surviving from one asexual generation to the next is $P \times (1 + s_h)$, where P is the probability that a parasite survives and undergoes a given round of replication and s_h is selection coefficient of the allele within the human host. Whether or not a parasite does or does not undergo a round of asexual reproduction is determined by the outcome of a Bernoulli trial. Each round of parasite replication creates two daughter cells. The maximum population size within a single human host is N_{eH}. After the population size reaches the maximum, it stays at the same population size for e additional WF generations.

(iii) The number of mosquitoes that obtain parasites (gametocytes) from each human host is based on multinomial sampling. If the mutation has a transmission advantage, the human host with the mutation has a $(1 + t_m)$-fold higher probability of transmitting gametocytes to mosquitoes. D

parasites (gametocytes) are transmitted from the human host to the mosquito host during each bite.

(iv) Within a mosquito host, the probability of surviving from one generation to the next is $P \times (1 + s_m)$, where s_m is the selection coefficient of the allele within the mosquito host. As in step (ii) in the human host, whether or not a parasite survives and undergoes a round of replication is determined by the outcomes of a Bernoulli trial, and if parasites do reproduce, they create two daughter cells. The maximum population size within the mosquito host is N_{eM}. The population size increases until it reaches maximum size.

(v) The number of humans that acquire parasites (sporozoites) from each mosquito host is based on multinomial sampling. If the mutation has a transmission advantage, the mosquito host with the mutation has a [(1+ t_h)-fold] higher probability to transmit sporozoites to human individuals.

(vi) Repeat steps (ii) to steps (v) until the mutation becomes lost or fixed in the entire population. Steps (ii) to (v) correspond to a “generation” in the malaria life cycle model.

We repeated the simulation 500,000 times for each initial condition. Table S4 lists all the relevant parameters and their default values. Unless stated otherwise, the default values were used in the simulations. Note especially that, while each human host can transmit parasites to multiple uninfected mosquito vectors, and each mosquito vector can infect multiple uninfected human hosts,

the model does not allow for any human host to be multiply infected by different parasite lineages. In other words, the simulation model is one of complete inbreeding. We chose complete inbreeding as the default model to mimic populations as in Senegal, where most infections are of single parasite genotypes (43). When selection takes place in the host, mixed infections would make selection even more efficient, and therefore our finding of enhanced efficiency of purifying selection would be even stronger if mixed infections were allowed. Simulations were performed using custom code written in C. This is available from the authors by request.

The results of the WF models used for comparison were also obtained by simulations. In the malaria model, we treated one complete life cycle as one generation and compared it with one generation in the WF model. We used 10,000 as the population size in the WF model because the default total parasite population size is 10,000 (10 transmitted parasites per host $\times$ 1000 hosts). The results of the WF models with population sizes 10^3, 10^8 and 10^{11} are also shown in Fig. S6. They are not qualitatively different from those obtained assume a population size of 10^4 and hence do not significantly alter the results of comparing the WF model with the malaria model.

Site-frequency spectrum

To obtain the null site-frequency spectrum that takes into account the malaria life cycle, we simulated mutations and kept track of them until they

543 became fixed or lost in the population. Then we sampled mutations weighted by
544 the time that they remained segregating in the population. Mutations that remain
545 in the population longer are more likely to be sampled. After a mutation was
546 chosen, we randomly selected one time point during the interval that the mutation
547 was segregating in the population, and at that time point sampled 159 parasites
548 from 159 different human hosts, and recorded the allele frequency. To minimize
549 the computational time for simulating new mutations, we first calculated the
550 relative probabilities of different initial conditions, and combined the site
551 frequency spectra according to their weighted average. Initial conditions with
552 probability less than 1/1000 of the most probable conditions contribute little to the
553 null distribution and therefore were ignored in the analysis (see Supplementary
554 Method for more details.).

ACKNOWLEDGMENTS

We thank John Wakeley, Russell Corbett-Detig, Stephen Schaffner, and Danny Milner for their helpful discussion and suggestions. We thank Ananias Escalante, Matt Berriman, James Cotton, and Yuseob Kim for reviewing the MS and providing the valuable comments that greatly improved the MS. This study is supported by NIH grant AI099105 as well as by grants from the Bill and Melinda Gates Foundation, the ExxonMobil Foundation, and the NIH Fogarty International Center.

References

1. Mu J*, et al.* (2010) Plasmodium falciparum genome-wide scans for positive selection, recombination hot spots and resistance to antimalarial drugs. *Nat Genet* 42(3):268-271.
2. Van Tyne D*, et al.* (2011) Identification and functional validation of the novel antimalarial resistance locus PF10_0355 in Plasmodium falciparum. *PLoS Genet* 7(4):e1001383.
3. Amambua-Ngwa A*, et al.* (2012) Population genomic scan for candidate signatures of balancing selection to guide antigen characterization in malaria parasites. *PLoS Genet* 8(11):e1002992.
4. Chang HH*, et al.* (2012) Genomic sequencing of Plasmodium falciparum malaria parasites from Senegal reveals the demographic history of the population. *Mol Biol Evol* 29(11):3427-3439.
5. Manske M*, et al.* (2012) Analysis of Plasmodium falciparum diversity in natural infections by deep sequencing. *Nature* 487(7407):375-379.
6. Park DJ*, et al.* (2012) Sequence-based association and selection scans identify drug resistance loci in the Plasmodium falciparum malaria parasite. *Proc Natl Acad Sci U S A* 109(32):13052-13057.
7. Lozovsky ER*, et al.* (2009) Stepwise acquisition of pyrimethamine resistance in the malaria parasite. *Proc Natl Acad Sci U S A* 106(29):12025-12030.
8. Brown KM*, et al.* (2010) Compensatory mutations restore fitness during the evolution of dihydrofolate reductase. *Mol Biol Evol* 27(12):2682-2690.
9. Costanzo MS, Brown KM, & Hartl DL (2011) Fitness trade-offs in the evolution of dihydrofolate reductase and drug resistance in Plasmodium falciparum. *PloS one* 6(5):e19636.
10. Toprak E*, et al.* (2012) Evolutionary paths to antibiotic resistance under dynamically sustained drug selection. *Nat Genet* 44(1):101-105.

11. Kappe SH, Vaughan AM, Boddey JA, & Cowman AF (2010) That was then but this is now: malaria research in the time of an eradication agenda. *Science* 328(5980):862-866.
12. Wright S (1931) Evolution in Mendelian Populations. *Genetics* 16(2):97-159.
13. Fisher RA (1930) *The genetical theory of natural selection* (Clarendon Press, Oxford).
14. Langley CH*, et al.* (2012) Genomic variation in natural populations of Drosophila melanogaster. *Genetics* 192(2):533-598.
15. Torgerson DG*, et al.* (2009) Evolutionary processes acting on candidate cis-regulatory regions in humans inferred from patterns of polymorphism and divergence. *PLoS Genet* 5(8):e1000592.
16. Fujimoto A*, et al.* (2010) Whole-genome sequencing and comprehensive variant analysis of a Japanese individual using massively parallel sequencing. *Nat Genet* 42(11):931-936.
17. Li Y*, et al.* (2010) Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. *Nat Genet* 42(11):969-972.
18. Carneiro M*, et al.* (2012) Evidence for widespread positive and purifying selection across the European rabbit (Oryctolagus cuniculus) genome. *Mol Biol Evol* 29(7):1837-1849.
19. Lee YC & Reinhardt JA (2012) Widespread polymorphism in the positions of stop codons in Drosophila melanogaster. *Genome Biol Evol* 4(4):533-549.
20. Slotte T, Foxe JP, Hazzouri KM, & Wright SI (2010) Genome-wide evidence for efficient positive and purifying selection in Capsella grandiflora, a plant species with a large effective population size. *Mol Biol Evol* 27(8):1813-1821.
21. Neafsey DE*, et al.* (2012) The malaria parasite Plasmodium vivax exhibits greater genetic diversity than Plasmodium falciparum. *Nat Genet* 44(9):1046-1050.
22. Der R, Epstein C, & Plotkin JB (2012) Dynamics of neutral and selected alleles when the offspring distribution is skewed. *Genetics* 191(4):1331-1344.
23. Otto SP & Whitlock MC (1997) The probability of fixation in populations of changing size. *Genetics* 146(2):723-733.
24. Hastings IM (1997) A model for the origins and spread of drug-resistant malaria. *Parasitology* 115 (Pt 2):133-141.
25. Hastings IM (2006) Complex dynamics and stability of resistance to antimalarial drugs. *Parasitology* 132(Pt 5):615-624.
26. Antao T & Hastings IM (2011) ogaraK: a population genetics simulator for malaria. *Bioinformatics* 27(9):1335-1336.

27. Schneider KA & Kim Y (2010) An analytical model for genetic hitchhiking in the evolution of antimalarial drug resistance. *Theoretical population biology* 78(2):93-108.
28. Kimura M (1955) Stochastic processes and distribution of gene frequencies under natural selection. *Cold Spring Harbor symposia on quantitative biology* 20:33-53.
29. Talman AM, Domarle O, McKenzie FE, Ariey F, & Robert V (2004) Gametocytogenesis: the puberty of Plasmodium falciparum. *Malar J* 3:24.
30. Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123(3):585-595.
31. Fay JC & Wu CI (2000) Hitchhiking under positive Darwinian selection. *Genetics* 155(3):1405-1413.
32. Fu YX & Li WH (1993) Statistical tests of neutrality of mutations. *Genetics* 133(3):693-709.
33. Charlesworth B, Morgan MT, & Charlesworth D (1993) The effect of deleterious mutations on neutral molecular variation. *Genetics* 134(4):1289-1303.
34. Begun DJ & Aquadro CF (1992) Levels of naturally occurring DNA polymorphism correlate with recombination rates in D. melanogaster. *Nature* 356(6369):519-520.
35. Smith JM & Haigh J (1974) The hitch-hiking effect of a favourable gene. *Genetical research* 23(1):23-35.
36. Hill WG & Robertson A (1966) The effect of linkage on limits to artificial selection. *Genetical research* 8(3):269-294.
37. Hedrick PW (1980) Hitchhiking: a comparison of linkage and partial selfing. *Genetics* 94(3):791-808.
38. Gardner MJ*, et al.* (2002) Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature* 419(6906):498-511.
39. Li H & Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26(5):589-595.
40. Li H*, et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25(16):2078-2079.
41. DePristo MA*, et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 43(5):491-498.
42. Volkman SK*, et al.* (2007) A genome-wide map of diversity in Plasmodium falciparum. *Nat Genet* 39(1):113-119.
43. Daniels R*, et al.* (2013) Genetic surveillance detects both clonal and epidemic transmission of malaria following enhanced intervention in Senegal. *PloS one* 8(4):e60780.

Figure legends

677 **Fig. 1. Synonymous and nonsynonymous site-frequency spectra.**
678 Synonymous and nonsynonymous site-frequency spectra are very similar (Mann-
679 Whitney test, *P* value = 0.46; *n* for synonymous and nonsynonymous sites are
680 4091 and 8778, respectively). Singleton SNPs were excluded in this analysis to
681 reduce the effect of sequencing errors.
682
683 **Fig. 2. Simulation diagram.** The simulation model is complete inbreeding
684 because it does not allow for any human host to be multiply infected by parasite
685 lineages from different mosquito vectors. The definitions of parameters are
686 shown in the main text and Table S4.
687
688 **Fig. 3. Comparison of probability of loss and average number of mutations**
689 **after one generation between the malaria life cycle and the WF model. (A)**
690 Probability of loss is greater in the malaria model. **(B)** Average number of
691 mutations after one generation is more extreme in the malaria model except
692 when the mutation is neutral.
693
694 **Fig. 4. Comparison of longer time scale properties between the malaria life**
695 **cycle and the WF model. (A)** Segregation time is shorter in the malaria model.
696 **(B)** Time to fixation for beneficial alleles is shorter in the malaria model when the
697 selection coefficient is smaller than threshold value ($s = 0.01$ under the default
698 settings), and is greater than in the WF model if the selection coefficient exceeds

699 the threshold. **(C)** Probability of fixation of beneficial alleles in the malaria model
700 is smaller than in the WF model, due to greater effects of random genetic drift
701 and stochastic transmission among hosts in the malaria life cycle. **(D)** Time to
702 loss of deleterious alleles is shorter in the malaria model, suggesting highly
703 efficient purifying selection in the malaria parasite.
704
705 **Fig. 5. Simulations for non-neutral transmissions.** In the "s_h-only" model (red
706 line), only s_h varies and transmission probabilities are all the same. In the "t_m-
707 only" model (yellow line), only t_m varies and s_h are all zero. In the "$t_m = s_h$" model
708 (blue line), t_m and s_h are the same. **(A)** In the t_m-only model, beneficial mutations
709 have shorter segregation times and deleterious mutations have slightly longer
710 segregation times than in the s_h-only model. **(B)** Fixation times for beneficial
711 mutations in the t_m-only model and the $t_m = s_h$ model are both shorter than in the
712 s_h-only model as well as in the WF model. **(C)** The fixation probabilities for
713 beneficial mutations in the t_m-only model and the $t_m = s_h$ model are larger than in
714 the s_h-only model, but still smaller than in the WF model. **(D)** The time to loss for
715 deleterious mutations is shorter in the s_h-only model and the $t_m = s_h$ model than in
716 the t_m-only model.
717
718 **Fig. 6. Allele-frequency spectrum of neutral alleles in the malaria model**
719 **compared with WF.** The allele-frequency spectrum in the malaria model is more
720 skewed toward lower frequency alleles than in the WF model.
721